Therapie (2019) 74, 181—185

Available online at
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

PHARMACOEPIDEMIOLOGY

The limitations of observation studies for decision making regarding drugs efficacy and safety

François Gueyffier[a,b,*], Michel Cucherat[a,c]

[a] CNRS, UMR5558, département biostatistiques et modélisation pour la santé et l'environnement, équipe évaluation et modélisation des effets des médicaments, laboratoire de biométrie et biologie évolutive, université Lyon 1, 69008 Lyon, France
[b] Service des données de santé, hospices civils de Lyon, 69003 Lyon, France
[c] Service de pharmacologie et de toxicologie, hospices civils de Lyon, 69003 Lyon, France

Received 3 September 2018; accepted 30 October 2018
Available online 8 November 2018

KEYWORDS
Pharmacoepidemiology;
Observational studies;
Level of evidence;
Biases;
Net benefit

Summary After looking at some case studies, we will remind in this paper how observational studies regarding drug exposure or delivery are prone to various biases and structural limitations. These biases lead us to be extremely cautious regarding the implementation of results from such studies in clinical practice, and to question the reliability of such studies to determine the position of a given treatment into the therapeutic strategy. We will conclude on the respective place of randomized controlled trials (RCTs) and observational approaches, which are obviously complementary, but not interchangeable.

© 2018 Société française de pharmacologie et de thérapeutique. Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations

COPD	chronic obstructive pulmonary disease
IPTW	inverse probability of treatment weighting
RCTs	randomized controlled trials

Introduction

Prescribing a given drug to a given patient is a common, but heavy decision. It relies on various elements among which the prediction of benefits of the drug exposure, expected to be reasonably higher than the risks of this exposure, beyond any reasonable doubt. This prediction is best based on the results from randomized clinical trials (RCTs) [1] and their meta-analyses, which answer three questions: (i) does a benefit exist from drug exposure in general?; (ii) is this benefit quantification enough to justify the drug exposure, in general, but also in this particular patient?;

* Corresponding author. UMR5558, unité de pharmacologie clinique et essais thérapeutiques, faculté de médecine Laennec, 7, rue G. Paradin, 69008 Lyon, France.
E-mail address: francois.gueyffier@univ-lyon1.fr (F. Gueyffier).

https://doi.org/10.1016/j.therap.2018.11.001
0040-5957/© 2018 Société française de pharmacologie et de thérapeutique. Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

(iii) are there any side effects, which are likely enough to compromise the positivity of the net benefit or the risk-to-benefit ratio, specifically expected for this patient situation? These answers are given to the community with the highest level of evidence, thanks to the reduction of biases through the methods of RCTs, ideally cumulating the most important items among the following: prospective design, with an a priori definition of hypotheses to be tested, an appropriate randomization, a double blind administration, an intention-to-treat analysis in agreement with the pre-specified analysis plan, and an appropriate hierarchy of tests to be conducted. Every distance taken to these standards reduce the methodological power of RCTs and its ability to control the impact of biases. Of course, some poorly designed or badly performed or badly analyzed RCTs had led to biased results [2]. The level of transparency and control by regulatory agencies could be improved, e.g. in conducting systematic re-analyses of principal results and allowing an efficient individual patient data sharing. And, for non-methodological reasons, RCTs are prone to over-select patients and, therefore, draw conclusion on only a subset of patients who are to treat in the real life [3] at the time of marketing authorization.

The medico-scientific community often considers that clinical trials are not appropriate to estimate rare adverse effects, or those resulting from long-term exposure to the drug. This consideration is logical given first, the size of samples studied in clinical trials, rarely above 10,000 patients and often less than 1000, and second, the duration of exposure, rarely above 5 years and often less than 1 year. These limitations are the main rationale to conduct observational studies of drug exposure consequences, to assess whether it could be associated with rare or delayed adverse events. Another important rationale of these studies is to observe drug impact in patients usually excluded from, or poorly represented in, RCTs samples. The third rationale for such studies is to describe the population to which the drug is prescribed after market access, in particular whether the administrative authorization constraints are respected. The growing availability of medical information in administrative care-providing databases, the development of analytical techniques for giant databases among which some invoke artificial intelligence, make observational approaches even more appealing. We will now look at some case studies to illustrate the weaknesses of observational studies regarding biases.

Introductory case studies

Could observational studies replace RCT to assess efficacy and safety of new treatment? Could observational studies be useful to confirm or infirm the results of the previous treatment's RCTs?

The limitations of observational studies to make causal inference are well known [1,4]. There are now multiple examples of RCT that did not confirm the efficacy of treatment, even showed by several observational studies. For example, the STATCOPE trial [5] failed to show the benefit of simvastatin for the prevention of exacerbations in moderate-to-severe chronic obstructive pulmonary disease (COPD), whereas several retrospective studies suggested a reduced rate of exacerbations in COPD patients receiving statins. However, these observational observations had appeared sufficiently reliable to ground the initiation of STATCOPE.

More recently, the ASCEND trial [6], comparing aspirin to placebo for primary prevention in people with diabetes mellitus, did not find benefit in terms of prevention of digestive cancer despite a large number of previous observational studies and meta-analysis which led to believe that aspirin had a marked preventing effect on digestive cancer.

These cases show that observational studies could lead to make a wrong decision about benefit or safety of drugs and could be of limited interested in decision-making. In fact, these restrictions arise from a series of limits: confounding, selection and information bias, loss to follow-up, reverse causation, uncontrolled statistical error risk, and purely inferential reasoning.

In the history of drug evaluation [7], these limitations were perceived early, and a lot of solutions were elaborated to put the clinical trials away from these problems. These solutions shaped the present design and methodology we use now with confidence to evaluate the efficacy and safety of news drugs: contemporary, randomized controlled clinical trial.

Confounding

Confounding is certainly one of the main difficulties of the observational setting. In real life, the decision to treat a patient with a drug is not taken at random but as a function of multiple variables like patient characteristics, severity of illness, socio-economic setting, etc. Among all these factors, some of them could be also linked to the outcome and will confound the association observed in the study between the treatment and the outcome.

To remove any confounding from an observational result and produce an appropriate picture of the true association, it should be necessary to consider all the confounding factors. In practice this is not achievable. Firstly, because it is difficult to identify all the factors that can act as confounders (by being linked both to the outcome and to the treatment study). Causal diagrams [8] are certainly the best option to conduct this analysis. Secondly, because in general, all potential confounders were not measurable directly or indirectly (especially with the claims database or all types of retrospective data). The question of unmeasurable confounders and residual confounding is crucial. Fewell et al. have shown that magnitude of the effect commonly estimated in the observational studies are of the same order of magnitude that the residual confounding generated by the unmeasured confounding factor [9].

In critical assessment of a result, before decision making, the use of a priori reasoning using knowledge [10,11] and causal diagrams in order to identify all the potential confounders is certainly the best way to give confidence to the reader that all confounders were potentially taken into consideration. The ultimate confidence will be given when all these variables were available in the dataset of the study.

The adjustment on a large number of variables performed to limit the risk of forgotten real confounders could be unproductive given the possibility of colliders. Colliders

[12] are variables that are causally influenced by two or more other variables. Conditioning on a collider, whatever the method used, leads to observe a non-causal association between these two variables. There is a real danger in extensive non-reasoned adjustment. The causal graphs are helpful to identify this kind of variable too.

In critical assessment of observational results in a decision-making perspective, it should be necessary to exclude with confidence the possibility of a residual confounding.

None of the available methods used to take into account the confounding (adjustment, matching, restriction by the confounders themselves or a summary of them like a propensity score, or an approach based on instrumental variables) could give the assurance of absence of residual confounding. In fact, it is not a problem of statistical method choice. Residual confounding depends only on the possibility to adjust on all the confounders [13]. In practice, it proofs impossible to be sure that analysis can adjust for all confounders. Discarding a residual confounding need another approach like the use of falsification variable or negative control [14—17]. Residual confounding could be ruled out if no association were detected between a set of appropriate falsification variable and the treatment under consideration. Falsification testing uses variables that are unlikely to be associated with the outcome but can be affected by confounding in the same way as the variable of interest. So in the absence of any residual confounding, no association should be observed between falsification variables and outcomes. These falsification variables are an elegant potential solution to validate findings from observational studies. The limit of this proof by *reductio ad absurdum* is the impossibility of being sure that all necessary falsification variables were used.

The sensitivity of a result to unmeasured confounders can also be assessed by different methods of sensitivity analysis [18,19].

Convincing positive proofs of absence of residual confounding is compulsory before actioning an observational result in decision-making. The description of a sophisticated adjustment method performed on a reasonable list of confounders is not enough to establish the reliability of the results. Applying a convincing method of diagnostic for residual confounding is therefore certainly mandatory to accept the result of an observational in decision-making.

Bias

Observational studies are also prone to all kinds of biases, independently to the confounding. The risk of bias in non-randomized studies of interventions I (ROBINS—I) ROB tools distinguishes 6 domains of bias:

- bias in selection of participants into the study;
- bias in classification of interventions;
- bias due to deviations from intended interventions;
- bias due to missing data;
- bias in measurement of outcomes;
- bias in selection of the reported result.

Given the diversity and the complexity of biases that potentially affect an observational study, it is virtually impossible to conclude that a result is not due, in totality or in part, to biases for weak to moderate association as frequently reported in observational pharmaco-epidemiological studies. Some principles were enacted to limit some biases, like the new users design, the validation of the proxy or generated variables in claims database, etc. but we do not yet have the demonstration that these principles lead to a full control of all the biases.

The bias analysis is even dependent of the nature of the results given that a null effect could be biased by a different kind of measurement error that a non-null effect (symmetric and non-symmetric error).

Studies on retrospective data

Another limitation is the use of retrospective data by much observational studies. In this setting, fishing expeditions or cheery picking are possible practices, leading to a dramatically increased risk of false discoveries. The prevention of these errors goes through a strict respect of the deducting reasoning with analyses conducted only after the definition of a precise objective in a pre-specified protocol, objectively stated without any knowledge of the result, which will be obtained. Mandatory protocol registering is part of the solution [20] but will never give absolute guarantees that the objective of the present protocol was not derived from a previous exploratory analysis on the same data [21].

Critical appraisal

Another difficulty arising when observational studies are considered for decision-making is the critical appraisal of the results. Critical appraisal of RCTs is performed mostly by checking if the study is protected by design against the bias. On the opposite in an observational study, this appraisal needs to determine if biases are present or not. As biased results are clearly undetectable by themselves (given that the truth is unknown), this exercise consists of a series of ''what if'' reasoning. It is not rare to obtain completely discordant appreciation of bias for the same study, each conclusion being based on a coherent reasoning.

No doubt that the recent availability of a ROBINS-I risk of bias assessment tool [22] will clarify and standardized these critical assessments.

Statistical considerations

Observational studies are also subject to high risk of type 1 error consecutively to uncontrolled multiplicity in the statistical comparisons. The notion of primary outcome is rarely used in these studies, and, in any event, is of little effectiveness in all retrospective studies.

Moreover, the adjustment-variables set is rarely predefined, leaving the possibility of some p hacking [23,24]. As mentioned by Rubin [25], in regression modeling there is always the temptation to work toward the desired or anticipated result.

To some extent, the vibration of results that could be generated by the adjustment choices is sufficiently large,

in general, to produce opposite results [26] and explain the magnitude of heterogeneity observed between observational studies on the same topic.

For example, in 2009, Diabetologia published a paper concluding to an increased risk of cancer with insulin glargine [27]. This purely exploratory study searched an association between insulin glargine and all a variety of cancers or aggregate of cancers. All comparisons were non-significant except for breast cancer. After this first paper, several other studies found similar results on all cancers, opening a controversy that could have last many years. In 2012, the ORIGIN [28] randomized controlled trial, comparing insulin glargine versus standard care in 12,537 people, ended the discussions by showing no effect on the pre-specified outcome of cancer (hazard ratio, 1.00; 95% CI, 0.88 to 1.13; $P = 0.97$).

''Real life'' data

The potential interest of the observational studies frequently spotlighted is their aptitude to estimate the drug effect in the real world, by considering more complex or frail patients than those enrolled in the RCTs. Let us remark that the ''real life'' designation is not appropriate, since randomized controlled trials recruit patients who are really living, and to whom we are all infinitely indebted for giving part of their life to improve our common knowledge bases.

However, some evidence [29] emerges that these complex and frail patients could not even be considered in the observational setting, due to several phenomena.

It turned out that doctors follow quite strictly the drug indications that are directly derived from the inclusion criteria of the pivotal phase 3. In the real world, the treated patients could be in some case very close to those studied in the RCTS [29].

On another hand, the methods used to take into account the confounders may lead to exclude patients of the analysis, for example, when trimming is performed in propensity score based analysis. This trimming, needed to ensure an optimal adjustment, leads to exclude the patients with the more extreme propensity scores, who represent the most atypical and complex patients.

In the popular method of matching, the control patients are matched to the patients treated with the drug under investigation. This method gives an estimation of the average treatment effect for the treated [30,31] and not for the overall population. To estimate marginal (or population-average) treatment effect, stratification or inverse probability of treatment weighting (IPTW) are needed [31].

Publication bias and selective reporting of outcomes

The reliability of observational study for decision-making is also hampered by an important risk of publication bias, or selective reporting of the results inside the publication. Studies on retrospective data are quite easy to perform and at a low-cost. The same association between treatment and an outcome can be repeatedly searched on several databases until a favorable result arise. Inside a same study, it can also be possible to repeat the search of the same association by varying the endpoint definition (almost all endpoints are derived from the aggregation of several items, ICD codes for example) or by testing several sub populations.

Conclusion

We have seen how observational studies are prone to a series of biases impossible to completely master reliably. This is why it is tremendously important to consider these studies and their results in their appropriate place. Their undisputable interest is to describe the exposed population, in particular whether prescriptions are in frontier zones of, or even in zones uncovered by, the marketing authorization. This information could be, per se, a signal important enough to question whether the benefit demonstrated in RCTs is really to be expected on such population, or in most members of this population.

Importantly, these studies cannot be trusted per se as a basis of clinical decision, either when they suggest that one drug is better or safer than another, or when they suggest an appropriate or conversely a deleterious risk to benefit ratio. Indeed, they are not able to inform correctly on such benefits or risks, even after multiple adjustment techniques application. Believing that these methods could replace randomization is non-sense, potentially misleading and source of erroneous pharmacological attitudes.

If they suggest a safety signal, this alarm should be considered with appropriate caution and not taken for granted before multiple checks, including estimates of residual confounding and exploring results on negative controls, ideally completed with ad hoc RCTs. Of course, in the case of an extraordinary safety signal, and in balancing with the size of the demonstrated benefit, appropriate decisions must be taken to prevent potential public health drama or crisis, and ideally trigger further confirmation studies including RCTs.

The heavy limitations of observational pharmacological studies must be acknowledged by medical community as a whole, they must be taught appropriately so that no confusion be made with the potential of causal conclusion that could be drawn from appropriately planned and conducted RCTs.

Our final point will be a hopeful one. The vast data warehouses that are more and more available in most countries will illustrate the limitations of observational approaches, in particular in showing the lack of replication for their confounded and biased results. But they will also (i) allow to better estimate risk factors impact, to better predict prognosis, thus to better adjust pharmacological treatment to appropriate target populations; and (ii) support the conduct of large and simple open randomized clinical trials of pharmacological or therapeutic strategies, in samples that will be more and more representative of the target or exposed population.

Disclosure of interest

The authors declare that they have no competing interest.

Take home messages

- The limitations of observation studies initiated the methods development of RCTs.
- Some RCTs conducted in the market access dossiers have practical limitations, mainly regarding duration and size.
- These limitations raise the question of a potential role of observation studies after market access.
- However, limitations of observation studies remain, and RCTs development, such as withdrawal or pragmatic trials, are the best response to market access dossiers constraints.
- Observational studies have an undisputable role in the definition of the joined population, which should develop greatly through administrative database.
- The complete inaptitude of observation studies at estimating benefit and security of therapeutics in a causal way must be acknowledged.

References

[1] Sackett DL. Why did the randomized clinical trial become the primary focus of my career? Value Health 2015;18(5):550—2.

[2] Le Noury F, Nardo JM, Healy D, Jureidini J, Raven M, Tufanaru C, et al. Restoring sutdy 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. BMJ 2015;351:h4320.

[3] Collet JP, Boissel JP. Sick population — treated population: the need for a better definition. The VALIDATA Group. Eur J Clin Pharmacol 1991;41(4):267—71.

[4] Beaglehole R, Bonita R, Kjellstrom T. Causation in epidemiology. Basic epidemiology. Geneva, Switzerland.: World health organization; 1993. p. 71—81.

[5] Criner GJ, Connett JE, Aaron SD, Albert RK, Bailey WC, Casaburi R, et al. Simvastatin for the prevention of exacerbations in moderate-to-severe COPD. N Engl J Med 2014;370(23):2201—10.

[6] ASCEND Study Collaborative GroupBowman L, Mafham M, Wallendszus K, Stevens W, Buck G, et al. Effects of aspirin for primary prevention in persons with diabetes mellitus. N Engl J Med 2018;379(16):1529—39, http://dx.doi.org/10.1056/NEJMoa1804988.

[7] Marks HM. The progress of experiment: science and therapeutic reform in the United States 1900—1990. Cambridge; New York: Cambridge University Press, Coll Cambridge history of medicine; 2000.

[8] Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB. Introduction to causal diagrams for confounder selection. Respirology 2014;19(3):303—11.

[9] Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol 2007;166(6):646—55.

[10] Hernán MA, Hernández-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155(2):176—84.

[11] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10(1):37—48.

[12] Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol 2010;39(2):417—20.

[13] Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, USA: Lippincott Williams & Wilkins; 1998.

[14] Arnold BF, Ercumen A, Benjamin-Chung J, Colford JM. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. Epidemiology 2016;27(5):637—41.

[15] Groenwold RHH. Falsification end points for observational studies. JAMA 2013;309(17):1769—70.

[16] Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21(3):383—8.

[17] Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? JAMA 2013;309(3):241—2.

[18] Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54(3):948—63.

[19] Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Ann Epidemiol 2008;18(8):637—46.

[20] Dal-Re R, Ioannidis JP, Bracken MB, Buffler PA, Chan AW, Franco EL, et al. Making prospective registration of observational research a reality. Sci Transl Med 2014;6(224):224cm1.

[21] Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: the ISPOR good research practices task force report. Value Health 2012;15(2):217—30.

[22] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.

[23] Bruns SB, Ioannidis JP. p-Curve and p-Hacking in Observational Research. PLoS One 2016;11(2):e0149144.

[24] Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 2014;15(1):1—12.

[25] Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services & Outcomes Research Methodology 2001;2:169—88.

[26] Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol 2015;68(9):1046—58.

[27] Jonasson JM, Ljung R, Talback M, Haglund B, Gudbjornsdottir S, Steineck G. Insulin glargine use and short-term incidence of malignancies-a population-based follow-up study in Sweden. Diabetologia 2009;52(9):1745—54.

[28] ORIGIN Trial investigatorsGerstein HC, Bosch J, Dagenais GR, Diaz R, Jung H, et al. Basal insulin and cardiovascular and other outcomes in dysglycemia. N Engl J Med 2012;367(4):319—28.

[29] Safieddine M, Ollier CC, Bertoletti E, Bellet L, Mismetti F, Cucherat P, et al. Comparison of randomized controlled trials and cohort studies for the assessment of direct oral anticoagulants (DOAC). Fund Clin Pharmacol 2018;32(Supp S1):12. https://onlinelibrary.wiley.com/doi/epdf/10.1111/fcp.12370 [Accessed 29 October 2018; Abstract CO-025].

[30] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46(3):399—424.

[31] Austin PC. A Tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivariate Behav Res 2011;46(1):119—51.